



# SPDI & NCBI Variation Data Processing Services

Processing variation data in various formats using the SPDI common data model

<https://api.ncbi.nlm.nih.gov/variation/v0>

National Center for Biotechnology Information • National Library of Medicine • National Institutes of Health • Department of Health and Human Services

## Scope and Access

The task of determining whether two genetic variants are the same poses certain challenges. We must navigate competing text file formats for variant representation, different standards for shifting ambiguous alignments, and deal with continually updated reference sequence models on which the variation calls are made. To help address these problems, we are publishing a set of versioned Variation Services for the genomic research community to use to group and compare variant observations. Variation Services use a common data model described as **Sequence Position Deletion Insertion** (SPDI, <https://www.ncbi.nlm.nih.gov/variation/notation/>). With these services, we are able to

- interconvert between HGVS and VCF formats, with proper left and right-shifting
- determine, if two variants are the same, using the same standard applied by ClinVar and dbSNP
- discover where the variant maps to on the current set of RefSeq sequence models
- retrieve RefSNP identifier by allele using VCF input, and
- retrieve detailed information for a RefSNP record as an JSON object.

For researchers with a one-off analysis need, you can refer to the API Documentation (below) as well as an NCBI blog-post (<https://go.usa.gov/xUuJc>) for tips on using the API. You can use this set of services as a functional replacement of the recently retired Variation Reporter (<https://go.usa.gov/xUJhf>). For software developers and workflow engineer, you can incorporate the services into your own data analysis pipelines, which would provide grouping of variants in the same way NCBI variation resources do it. Our initial release (version 0) and every major version thereafter will maintain the backward-compatibility of the object schema.

NIH U.S. National Library of Medicine  
National Center for Biotechnology Information

<https://api.ncbi.nlm.nih.gov/variation/v0/>

## Variation Services

Services for variation data processing

Visit the NCBI Variation Services documentation page to learn more about the functionality of these services.  
<https://www.ncbi.nlm.nih.gov/variation/services/>

Created by NCBI Variation Services Team  
[Contact the developer](#)

powered by Swagger

**SPDI : Services receiving input alleles as SPDI objects in colon-separated format**

Method	Endpoint	Description
GET	/spdi/{spdi}/contextual	Retrieve contextual allele in SPDI syntax for provided allele
GET	/spdi/{spdi}/vcf_fields	Retrieve fields suitable for representing the input allele in VCF
GET	/spdi/{spdi}/canonical_representative	Retrieve canonical allele representative
GET	/spdi/{spdi}/all_equivalent_contextual	Retrieve equivalent contextual allele on all aligned sequences
GET	/spdi/{spdi}/rsids	Lookup the RSIDs (if any) associated with the input allele
GET	/spdi/{spdi}/hgvs	Retrieve HGVS notation for contextual allele in SPDI Syntax

**HGVS : Services receiving input alleles in HGVS notation**

Method	Endpoint	Description
GET	/hgvs/{hgvs}/contextuals	Retrieve the contextual alleles equivalent to the HGVS notation input
POST	/hgvs/batch/contextual	Retrieve contextual alleles for each input

**VCF : Services receiving input alleles in VCF notation**

Method	Endpoint	Description
GET	/vcf/{chrom}/{pos}/{ref}/{alts}/contextuals	Retrieve contextual alleles for the reference allele and each alternate allele
POST	/vcf/file/set_rsids	For provided VCF data, add RSIDs in the ID field

**RefSNP : Services accessing RefSNP properties**

Method	Endpoint	Description
GET	/beta/refsnp/{rsid}	Lookup RefSNP object by RSID

[ BASE URL : /variation/v0 , API VERSION : 0.1.3 ]

Log in

## Accessing the Variation Services

As a start, you can access the Variation Services through its landing page shown to the left (A). The top of this page (B) provides a summary description and a link to the help document with additional details. The body of the page groups functions available into major categories, each under their own heading:

- The SPDI section (C) contains functions that use SPDI as input and returns output in other format such as VCF, HGVS, and RSID.
- The HGVS section (D) contains functions that use HGVS as input and return the variant context in SPDI format.
- The VCF section (E) contains functions that takes VCF input and returns either the SPDI context or rsids.
- The last RefSNP section (F) contains a function that takes an RSID as input and returns a JSON data blob similar to the content used for the newly redesigned RefSNP page, such as <https://www.ncbi.nlm.nih.gov/snp/rs268>.

Under each heading, available functions appear in separate lines, listing the base URL along with a phrase describing that function.

## Accessing the Variation Services (cont.)

The Variation Services landing page displays each available functions by:

- showing the operation mode as a button, with **GET (A)** in blue and **POST** in green (POST mode accepts batch input)
- listing its base URL construct **(B)**, and
- describing the task the function accomplishes **(C)**

You can click the button to toggle open the display of a function and see example input in the Parameters subsection **(D)**, click “Try it out!” button **(E)** to test the service with sample input, and see the actual command line as well as the actual output.

## Usage Examples for Functions Taking HGVS and VCF as Input

The third and fourth group of functions provided by the Variation Services take HGVS and VCF as input, respectively.

### 1. HGVS : Services receiving input alleles in HGVS notation

#### 1a. For single HGVS expression

As the description stated, this service takes variant in HGVS notation **(D)**, and retrieves the contextual alleles (in JSON format, **F**). For HGVS without explicit reference to a specific Reference Sequence record (accession. Version), you will need to provide the assembly accession. Version **(G)**. If you do not know this for the specific assembly you are using to make your variant calls, you can find it out by searching in the [NCBI Assembly database](https://www.ncbi.nlm.nih.gov/assembly/).

Clicking “Try it out!” button **(E)**, Variation Services will execute with the provided input and append the result in an added section below. The yellow textboxes **(H)** from the top to bottom provide

- the curl command line with its parameter/value pairs
- the URL for the request, with the unsafe characters escaped
- the services' response in JSON format
- the status code, and
- the response header

You can model the curl command line or the request URL, and incorporate them in your own workflow need.

**HGVS : Services receiving input alleles in HGVS notation** [Show/Hide](#) [List Operations](#) [Expand Operations](#)

**GET** **/hgvs/{hgvs}/contextuals** **Retrieve the contextual alleles equivalent to the HGVS notation input**

**Implementation Notes**  
For the input HGVS notation, retrieve all contextual alleles in SPDI syntax. (There can be more than one because of differences in the way the two models represent variation. For example, NC\_012920.1:m.961delTinsC(2\_7) requires one SPDI for each number of cytosines inserted.)

**Response Class (Status 200)**

**Model | Model Schema**

```
{
  "data": {
    "spdis": [
      {
        "seq_id": "NC_000001.23",

```

**Response Content Type**  
application/json

**Parameters**

Parameter	Value	Description	Parameter Type	Data Type
hgvs	NC_000001.10:g.12345T>A	Allele in HGVS notation	path	string
assembly	GCF_000001405.25	GenColl assembly accession for provided list of HGVS. Required only if the hgvs uses a chr location like chr1:g.12345A>T or chrX:g.234C>G	query	string

**Response Messages**

HTTP Status Code	Reason	Response Model	Headers
default	Unexpected error	Model   Model Schema	

```
{
  "error": {
    "code": 0,
    "message": "string",
    "errors": [

```

**Try it out!**

**Curl**

```
curl -X GET --header "Accept: application/json" "https://api.ncbi.nlm.nih.gov/variation/v0/hgvs/NC_000001.10%3Ag.12345T%3EA/contextuals?assembly=GCF_000001405.25"
```

**Request URL**

```
https://api.ncbi.nlm.nih.gov/variation/v0/hgvs/NC_000001.10%3Ag.12345T%3EA/contextuals?assembly=GCF_000001405.25
```

**Response Body**

```
{
  "data": {
    "spdis": [
      {
        "seq_id": "NC_000001.10",
        "position": 12344,
        "deleted_sequence": "T",
        "inserted_sequence": "A"
      }
    ]
  }
}
```

**Response Code**  
200

**Response Headers**

```
{
  "ncbi-sid": "EC23F7C8B72D7561_0000SID",
  "date": "Mon, 27 Aug 2018 18:09:56 GMT",
  "content-encoding": "gzip",
  "status": "200",

```

## 1. HGVS : Services receiving input alleles in HGVS notation (cont.)

### 1b. For batch HGVS expressions

This section does the same general task, providing the contextual information for input variant expressed in HGVS format, but use POST mode allows it to take a batch of HGVS input up to 50,000 expressions (A). Clicking "Try it out!" button (B), Variation Services will execute with the provided input and append the result in an added section below. The yellow textboxes (C) from the top to bottom provide

- the curl command line with its parameter/value pairs
  - the base-URL for the request (the JSON input is submitted through POST is not included in the URL)
  - the services' response in JSON format
- The status code and the response header boxes (similar to 1a) are not shown.

**B** Try it out! Hide Response

Curl

```
curl -X POST --header "Content-Type: application/json" --header "Accept: application/json"
```

Request URL

[https://api.ncbi.nlm.nih.gov/variation/v0/hgvs/batch/contextuals?assembly=GCF\\_000001405](https://api.ncbi.nlm.nih.gov/variation/v0/hgvs/batch/contextuals?assembly=GCF_000001405)

Response Body

```
{
  "data": [
    {
      "hgvs": "NC_000001.10:g.12345T>A",
      "alleles": {
        "spdis": [
          {
            "seq_id": "NC_000001.10",
            "position": 12344,
            "deleted_sequence": "T",
            "inserted_sequence": "A"
          }
        ]
      }
    }
  ]
}
```

**C**

**POST** /hgvs/batch/contextuals Retrieve contextual alleles for each input

Implementation Notes

For the input HGVS notation, retrieve all contextual alleles in SPDI syntax. (There can be more than one because of differences in the way the two models represent variation. For example, NC\_012920.1:m.961delTinsC(2\_7) requires one SPDI for each number of cytosines inserted).

Response Class (Status 200)

Model | Model Schema

```
{
  "data": [
    {
      "hgvs": "string",
      "alleles": {
        "spdis": [
          {
            "seq_id": "string",
            "position": "integer",
            "deleted_sequence": "string",
            "inserted_sequence": "string"
          }
        ]
      }
    }
  ]
}
```

Response Content Type

application/json

Parameters

Parameter	Value	Description	Parameter Type	Data Type
hgvs	<pre>{   "hgvs": [     "NC_000001.10:g.12345T&gt;A",     "NT_005612.16:g.36609556delC"   ] }</pre>	JSON object containing a single field "hgvs" which contains an array of the variants to process, each in HGVS notation. Up to 50,000 expressions may be included in a single request.	body	Array[string]

Parameter content type:

application/json

**A**

## 2. VCF : Services receiving input alleles in VCF notation

### 2a. For single variant in VCF format

This service (shown to the right) takes a variant in VCF notation provided through a set of input boxes (D). Refer to the Description column for details regarding the value you need to provide to each input fields. For input to chromosome field without explicit reference to a specific Reference Sequence record (accession.version), you will need to provide the assembly accession. You can obtain the information from the NCBI Assembly database.

Clicking the "Try it out!" button (E), the service will take the input, execute the request, and returns the contextual information in JSON format.

**GET** /vcf/{chrom}/{pos}/{ref}/{alts}/contextuals Retrieve contextual alleles for the reference allele and each alternate allele

Implementation Notes

Returns a list of SPDI format alleles containing one contextual allele for each reference and alternate allele specified by the input VCF fields.

Response Class (Status 200)

Model | Model Schema

```
{
  "data": {
    "spdis": [
      {
        "seq_id": "NC_000001.23",
        "position": 0,
        "deleted_sequence": "string",
        "inserted_sequence": "string"
      }
    ]
  }
}
```

Response Content Type

application/json

Parameters

Parameter	Value	Description	Parameter Type	Data Type
chrom	NC_000001.10	Usually this is the RefSeq/Genbank Accession.Version for the reference sequence. Despite the name (taken from the VCF standard) this does not need to be a chromosome. But this field can also be an integer 1..22 or a string. The string can be "X", "Y", or be of the form like "chr2". In those cases the assembly parameter must be supplied to tell from what assembly the chromosome comes.	path	string
pos	12345	The 1-based position on the reference sequence of the first nucleotide in the reference allele string	path	integer
ref	T	The reference allele, in IUPAC notation, with padding nucleotide when required.	path	string
alts	A,G	Comma delimited list of alternate alleles, in IUPAC notation, with padding nucleotide when required	path	string
assembly	GCF_000001405.25	GenColl accession.version string of the assembly used to disambiguate references. Required if the chrom field uses a location like '1' or 'chr1'	query	string

Response Messages

**E** Try it out!

[section removed for brevity]

**D**

## 2. VCF : Services receiving input alleles in VCF notation (cont.)

**Try it out!** **A**

Curl

```
curl -X GET --header "Accept: application/json" "https://api.ncbi.nlm.nih.gov/variation/v0/vcf/NC_000001.10/12345"
```

Request URL

https://api.ncbi.nlm.nih.gov/variation/v0/vcf/NC\_000001.10/12345/T/A%2CG/contextuals?assembly=GCF\_000001405.25

Response Body

```
{
  "data": {
    "spdis": [
      {
        "seq_id": "NC_000001.10",
        "position": 12344,
        "deleted_sequence": "T",
        "inserted_sequence": "A"
      },
      {
        "seq_id": "NC_000001.10",
        "position": 12344,
        "deleted_sequence": "T",
        "inserted_sequence": "G"
      }
    ]
  }
}
```

**B**

The service appends the result in a new section (**A**) below the button, which contains the curl command line and request URL, followed by actual server response (SPDI allele context in JSON format, **B**). The status code and response header fields are further below (not shown).

### 2b. For batch VCF-formatted input

The main purpose of this service is to map a set of input variants called in a study to existing entries in dbSNP. You can paste up to 50,00 lines of variants, expressed in 5-column VCF expression, in the input box (**C**). The service overwrites the custom identifiers in the 3rd column with matched RSIDs or NORSID if such mapping fails (**D**).

**POST** **/vcf/file/set\_rsids** For provided VCF data, add RSIDs in the ID field

**Implementation Notes**

Where there are matches to dbSNP, update the ID column of the VCF file with the matching RefSNP Identifiers (RSIDs), overwriting any pre-existing data.

**Response Class (Status 200)**

**Response Content Type**

text/plain; charset=utf-8

**Parameters**

Parameter	Value	Description	Parameter Type	Data Type
vcf_rows	NC_000010.11 190345 V10_1 C T NC_000001.11 1430245 V1_1 A G NC_000008.11 190298 V8_1 G T	Up to 50,000 VCF rows may be included in a single request. Only four fields will be used - CHROM, POS, REF and ALT	body	string
assembly	GCF_000001405.31	GenColl accession.version string of the assembly used to disambiguate references. Required if the chrom field uses a location like '1' or 'chr1'	query	string

**Response Messages**

*[section removed for brevity]*

**Try it out!** **C**

**Hide Response**

Curl

```
curl -X POST --header "Content-Type: text/plain; charset=utf-8" --header "Accept: text/plain; charset=utf-8" -d "NC_000010.11 190345 . C T  
NC_000001.11 1430245 V1_1 A G  
NC_000008.11 190298 V8_1 G T  
"https://api.ncbi.nlm.nih.gov/variation/v0/vcf/file/set_rsids?assembly=GCF_000001405.31"
```

Request URL

https://api.ncbi.nlm.nih.gov/variation/v0/vcf/file/set\_rsids?assembly=GCF\_000001405.31

**Response Body** **D**

NC_000010.11	190345	rs748857305	C	T
NC_000001.11	1430245	rs1326380815	A	G
NC_000008.11	190298	NORSID	G	T

### RefSNP : Services accessing RefSNP properties

**GET** **/beta/refsnp/{rsid}**

**Implementation Notes**

Retrieve data object associated with a RefSNP identifier for the service and the resulting schema is still under development.

**Response Class (Status 200)**

*[section removed for brevity]*

**Parameters**

Parameter	Value
rsid	328

**Response Messages**

*[section removed for brevity]*

**Try it out!** **E**

**Hide Response**

Curl

```
curl -X GET --header "Accept: application/json" "https://api.ncbi.nlm.nih.gov/variation/v0/beta/refsnp/328"
```

Request URL

https://api.ncbi.nlm.nih.gov/variation/v0/beta/refsnp/328

Response Body

```
{
  "refsnp_id": "328",
  "create_date": "2000-09-19T17:02Z",
  "last_update_date": "2018-05-11T06:02Z",
  "last_update_build_id": "151",
  "dbsnp1_merges": [
    {
      "merged_rsid": "3735962",
      "revision": "108",
    }
  ]
}
```

**F**

**G**

**H**

[ftp://ftp.ncbi.nih.gov/snp/redesign/latest\\_release/JSON](ftp://ftp.ncbi.nih.gov/snp/redesign/latest_release/JSON)  
<https://github.com/ncbi/dbsnp/tree/master/tutorials>

### 3. RefSNP : Services accessing RefSNP properties

The service under this section provides direct access to the complete SNP record in JSON format when given a valid RSID. The example (left) uses the RSID 328 as input (**E**) to retrieve the JSON object for that SNP record (**F**). You can readily incorporate the function in your own workflow by modeling after the command line or request URL examples (**G**). For the complete SNP dataset in JSON format, use the FTP site instead. Tutorials on how to process the JSON dataset is also available online (**H**).